Multi-echelon systems: A service measure perspective

E.B. Diks a, A.G. de Kok a,*, A.G. Lagodimos b

a Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
b Department of Business Administration, University of Piraeus, 80 Karaoli & Dimitriou Street, GR-185 34 Piraeus, Greece

Received 1 March 1995

Abstract

This paper reviews the most important results on divergent multi-echelon systems. In particular, we concentrate on the interactions between the elements that constitute such a multi-echelon system, in order to determine several service measures (e.g. external customer service level and inventory holding costs). We distinguish between two types of policies: installation stock and echelon stock policies. A comparison between these two types of policies revealed that the complexity of the analysis is concentrated at different aspects, which are discussed by reviewing the most important papers on both types. Special attention is given to the applicability of the models. Extensions to divergent multi-echelon systems with more than two stages are also treated.

Keywords: Inventory; Multi-echelon system; Periodic review; Continuous review; Service level

1. Introduction

Over the last decade considerable emphasis is put on the coordination of all operations of the material supply chain. Traditional logistics operations have taken place on the basis of buying material at a supplier and selling products to customers without explicit exchange of information other than prices and lead times. However, the decrease of product life cycles has unveiled the need for frequent exchange information on material availability and forecasts upstream in the supply chain as well as capacity information downstream on a routine basis. Experience on such information exchange has been gained within large vertically integrated companies during the seventies and eighties. It was found that a prerequisite for effective coordination of the supply chain is the measurement of operational performance in terms of due date reliability, stock availability and other customer service measures. This enables partners to update relevant information on customer demand and supply performance, so that effective supply chain management is achieved.

One of the main difficulties of cost-efficient and effective supply chain management is to determine the target service levels (associated with the service measures selected), so that the prespecified external service targets are met at minimum cost. Nowadays such targets are based on historical data and intuitive reasoning. The quantitative analysis of multi-echelon systems contributes to the solution of this problem, since it enables to predict the performance of a supply chain given the performance targets for individual stockpoints. Much progress on this analysis has been made in the last decade, as we will show in the sequel.
In this paper we review the most important results of three decades of multi-echelon theory from a service measure perspective. This paper complements the survey papers [70] and [28], which focus on cost-optimization issues. Since shortage costs are often hard to determine in practice, service measures are used as operational surrogates. Even in situations where shortage costs can be determined (e.g. in the case of contractual arrangements regarding shortage penalties), service measures are needed in order to have direct information on physical performance of the supply chain.

In this paper we distinguish between two different types of performance measures along the supply chain: internal and external performance measures. The latter are related to the service provided to external customers at the most downstream stockpoints of the supply chain (e.g. the fill rate and customer waiting times). The former are related to internal customer service (e.g. fill rate), and relevant costs (e.g. ordering, holding and transportation costs). We emphasize that, in some sense, internal service is irrelevant as long as external service is according to prespecified targets at a minimum internal cost.

A prerequisite to determine these performance measures is to have models to analyze the physical behavior (i.e., operational characteristics) of multi-echelon systems. Therefore we concentrate on analyzing the interactions between the elements that constitute a multi-echelon system. When such analysis yields insight into the evolution of material stocks of a multi-echelon system over time we are able to determine the service measures and the costs. Like in the companion review paper [70], we primarily focus on numerical tractability and applicability of these analyses, rather than analytic optimality.

The paper is organized as follows. In Section 2 we introduce the major elements that constitute a multi-echelon system and affect its control. In our review we use these elements as a classification instrument. For instance, we distinguish between so-called installation stock and echelon stock policies. In Section 3 we discuss the major results regarding installation stock policies. We distinguish between supply chains for consumable products and repairable products (spare parts). As much as possible we give a unifying treatment of subsequent contributions to show the progress made with regard to real-world systems. In Section 4 we give an extensive treatment of echelon stock policies. Special attention is given to the notion of imbalance, which is specific to echelon stock policies in the case of divergent supply chains. Another important aspect of echelon stock policies discussed in Section 4 is the rationing rule in situations where a stockpoint does not have sufficient stock to satisfy all downstream stockpoints. In installation stock policies rationing is not considered since one typically assumes FCFS. Based on our review of the literature in Sections 3 and 4 we propose directions for future research in Section 5.

2. Multi-echelon system elements

In the planning and control of a supply chain we distinguish between two kinds of network structures, which are the building blocks for more complex network structures. Usually the upstream part of a chain is characterized by a convergent structure. For instance, several components are assembled into one subassembly or finished product. Such an assembly stage may be subdivided in several phases separated by intermediate stockpoints. After the assembly stage the finished product is stored at a central depot, which supplies a number of downstream stockpoints. The distribution of such a finished product from the central depot to the end-stockpoints is characterized by a divergent structure.

For the analysis of convergent multi-echelon systems (e.g. assembly systems) we refer to [70]. We focus on the literature concerning the control of divergent multi-echelon systems. In Section 2.1 we present the divergent multi-echelon system under consideration. The behavior of the stock levels in such a system, depends on the ordering policies of the stockpoints. In Section 2.2 we present some practically useful continuous and periodic review ordering policies. For every ordering policy we distinguish between two variants: the installation stock policy and the echelon stock policy. In Section 2.3 we present the operating details of both variants, and demonstrate the differences. The order policy and its control parameters affect the internal customer service as well as the service provided to external customers. In Section 2.4 we define three major service measures.
We introduce some notation which will be used in the remainder of this paper:

\[ L_0 := \text{lead time from supplier to the central depot}, \]
\[ L_i := \text{lead time from central depot to retailer } i, \]
\[ M := \text{number of retailers}, \]
\[ \alpha_i := \text{the probability that the net stock (stock on hand minus back orders) at retailer } i \text{ is non-negative at the end of an arbitrary replenishment cycle}, \]
\[ \beta_i := \text{fraction of the demand satisfied directly from the stock on hand at retailer } i, \]
\[ \gamma_i := \text{one minus the ratio of the average shortage at retailer } i \text{ immediately before arrival of a replenishment order and the average demand at retailer } i \text{ during an arbitrary replenishment cycle}, \]
\[ D_{i,t+v} := \text{demand at retailer } i \text{ in } [t, t+v), \]
\[ D_{t,t+v} := \text{aggregate system demand in } [t, t+v); \]
\[ \delta_i := \text{the (echelon) inventory position of retailer } i \text{ at time } t \text{ just after ordering (rationing)}, \]
\[ S_i := \text{order-up-to level of stockpoint } i, \]
\[ s_i := \text{reorder-point of stockpoint } i, \]
\[ Q_i := \text{order quantity of stockpoint } i, \]

where stockpoint \( i \) refers to retailer \( i \) for \( 1 \leq i \leq M \) and refers to the depot for \( i = 0 \). In principle the variables \( L_0, L_i, D_{i,t+v}, D_{t,t+v} \) and \( \delta_i \) are random variables. Note that we use the notation for a two-echelon divergent system. In general for \( N \)-echelon models a different notation is required (cf. [72]), which we omit here for sake of clarity of the exposition.

2.1. Divergent systems

A divergent multi-echelon system is characterized by the property that a stockpoint is supplied by exactly one other stockpoint, and supplies one or more stockpoints. An \( N \)-echelon system is a multi-echelon system where the highest number of stockpoints on a path between the unique root stockpoint of the system and an end-stockpoint equals \( N \). Most papers in the literature restrict to two-echelon systems \((N = 2)\), in which the unique stockpoint, also called the depot, supplies \( M \) end-stockpoints, which are called retailers (see Fig. 1). Only these retailers face stochastic customer demand, which is stationary at each retailer, and independent of the demands at the other retailers. The supplier of the depot has an infinite capacity, i.e., whenever the depot places a replenishment order, this can be delivered after a lead time \( L_0 \). The lead time from the depot to retailer \( i \) is denoted by \( L_i \). Like in most papers we explain the analysis by considering this divergent two-echelon inventory system. However, when the results can be extended to the more general divergent \( N \)-echelon system this will be pointed out.

In this paper we concentrate on the specific problems occurring in divergent systems. Therefore we will not address the \( N \)-serial system in much detail. The \( N \)-serial system is a specific case of the divergent \( N \)-echelon system, in which every stockpoint has a unique supplier, but also has a unique successor. Quite some work in the literature has been devoted to this system. For an extensive overview on papers using dynamic programming we refer to the survey paper [27]. Another cost-related approach to analyze these serial systems is presented in [70]. Other more service related papers on this subject are [7,16,10,39,40].

2.2. Ordering policies

This paper addresses two major order disciplines: continuous review policies and periodic review policies. In the case of continuous review, the stock level is monitored constantly and, immediately after this level drops below a reorder point, an order can be placed to replenish the stock. Furthermore, the demand pro-
cess is usually assumed to be a compound Poisson process, or simply a Poisson process. In the case of periodic review, the stock level is inspected periodically, so that orders are generated at review moments only. In this case the demand per review period may have an arbitrary distribution.

In our review we focus on practically useful replenishment policies, which are easy to implement and do not necessarily have to be cost-optimal. Examples of such policies are the order-point, order-up-to policy and the order-point, order-quantity policy. The order-point, order-up-to policy is characterized by the control parameters \((s, S)\). This means that at any time when the stock level at equals or drops below \(s\), an order is placed immediately. The quantity of this order is such that the stock level returns to a target order-up-to level \(S\). This policy has been proven to be optimal for a single location inventory system which assumes a constant ordering cost, linear holding and stock-out costs, fixed replenishment lead time and backorder of unsatisfied demand \([33,34]\). Under some assumptions (among other things orders do not cross in time) Kaplan \([37]\) extends this result by proving that the \((s, S)\) policy is also optimal for the random replenishment lead time case. The periodic review analogue of the \((s, S)\) policy is denoted by \((R, s, S)\) (if \(s\) equals \(S\), then the \(s\) is suppressed). This means that every \(R\) time units the inventory is inspected, and orders are generated at these review moments only. The order-point, order-quantity policy is characterized by \((s, nQ)\). This means that when the stock level \(x\) falls to or below reorder-point \(s\) an order of \(nQ\) products is placed where \(Q\) is the base order-quantity and \(n\) is the minimum integer with \(x + nQ > s\). The periodic review analogue of the \((s, nQ)\) policy is denoted by \((R, s, nQ)\).

### 2.3. System control

The control of multi-echelon systems is often completely decentralized in the sense that ordering decisions at a stockpoint are solely based on the installation stock, i.e., the inventory position at this stockpoint. The inventory position is defined as the sum of all planned orders at this stockpoint and its physical stock minus its back orders. An obvious advantage of an installation stock policy is that it does not require any information about the inventory situation at other stockpoints. However, due to this lack of information about the entire system the cost effectiveness of these policies is limited. E.g. excessive demand may not be identified at upstream stockpoints due to the delay in information through resulting replenishment orders upstream. One way of taking such information into account is to control the inventory based on the echelon stock, i.e., the echelon inventory position at this stockpoint. The echelon inventory position is defined as the sum of all planned orders at this stockpoint plus its physical stock plus that in transit to or on hand at its downstream stockpoints minus eventual back orders at its end-stockpoints. Since in echelon stock policies the order decisions are based on the complete knowledge of how much stock is downstream, we need information on how the products flows through the system. Due to the developments in the area of information technology this is not a problem anymore.

The definition of the echelon inventory position given above can be seen as an analogue of the inventory position. However, it is more common to define the echelon inventory position of a stockpoint by all stock in transit to this stockpoint plus its physical stock plus that in transit to or on hand at its downstream stockpoints minus back orders at its end-stockpoints. To illustrate the possible difference between the two definitions, let us consider the case where a stockpoint uses a periodic review policy. In case all review moments coincide with arrival times of replenishment orders, then both definitions yield identical material flows. However, if replenishment orders arrive between two review moments then material flows may differ. If back orders at the supplying stockpoint are included in the echelon inventory position, then an arrival of a replenishment order at this supplying stockpoint between two subsequent review moments may lead to partial shipments to resolve this back-order position. This cannot occur when only in transit stock is included in the inventory position. In case of a continuous review policy, in principle, both definitions lead to the same material flow in the system (depending on how an arriving order is allocated to its successors). In the remainder of this paper we shall use the latter definition, just like most papers using the echelon stock concept.

An important difference between installation stock and echelon stock policies is pointed out by Chen and Zheng \([10]\). In the former policies the inventory po-
sition of a stockpoint includes all outstanding orders, i.e., in transit to this stockpoint or back-ordered at the supplier, while for echelon stock policies the echelon inventory position of a stockpoint only includes the in transit orders to this stockpoint. As a consequence, a stockpoint using an installation stock policy can always raise its inventory position to a desired level. If some part of the order cannot be delivered immediately it is back-ordered at its supplier. So we may model every stockpoint as a single location inventory system with a random lead time, i.e., the transportation time plus an additional waiting time. In echelon stock policies this lead time exactly equals the transportation time. However, in these policies it is more difficult to determine the echelon inventory position, since the stockpoint cannot be regarded as a single location inventory system.

Recently Axsäter and Rosling [5] compared installation stock policy and echelon stock policy. They proved that when every stockpoint in a multi-echelon system is controlled by an order-up-to policy an installation stock policy can always be replaced by an equivalent echelon stock policy, and vice versa. When every stockpoint in a serial system is controlled by an \((s, nQ)\) policy an installation stock policy can always be replaced by an equivalent echelon stock policy, but not vice versa.

2.4. Service measures

In order to determine good (or even optimal) control parameters of the ordering policy we need to determine the costs of holding inventory, order costs and the cost of stock-outs. As argued in [51] the stock-out cost may be ascribed to a lost sale or a rush delivery. More often however, the primary consideration is the possible loss of customer goodwill. One usually is not able, in practice, to assign these costs. Hence, they are determined indirectly by a certain service level. In this paper we consider three different service measures [38]:

- The non-stock-out probability \((\alpha)\): the probability that the net stock (the stock on-hand minus back orders) is non-negative at the end of an arbitrary replenishment cycle.
- The fill rate \((\beta)\): the fraction of the period demand that is satisfied directly from the stock on hand.
- The modified fill rate \((\gamma)\): one minus the ratio of the average shortage immediately before arrival of a replenishment order and the average demand during an arbitrary replenishment cycle.

In the determination of these service levels the difference between installation and echelon stock policies becomes clear. For instance, consider the \(\alpha\)-service level of retailer \(i\),

\[
\alpha_i = \Pr\{I_i^t - D_{t+L_i+W_t} \geq 0\},
\]

where \(W_t\) denotes an additional waiting time. In installation stock policies it is fairly easy to obtain an expression for \(I_i^t\) (e.g. in order-up-to policies \(I_i^t = S_t\)). However, the order placed at time \(t\) arrives after the transportation time \(L_i\) plus the additional waiting time \(W_t\). Usually it is hard to obtain the distribution of \(W_t\), since it generally depends on the parameters of the ordering policies and the characteristics of the demand processes. In echelon stock policies \(W_t\) simply equals zero. However, it is rather cumbersome to obtain an expression for \(I_i^t\), since in general it depends on the parameters of the ordering policies and the characteristics of the demand processes. In Section 3 we give some analytical expressions for the service levels defined for installation stock policies (e.g. the fill rate and the modified fill rate in Section 3.1.2), and in Section 4 we give expressions for the service levels for periodic review echelon order-up-to policies (e.g. the non-stock-out probability, the fill rate and the modified fill rate in Section 4.1.5).

3. Installation stock policies

In this section we review important contributions in controlling divergent multi-echelon systems using the installation stock concept. Most of the papers [21,30,64,32] determine the service performance of such a system for a given specification of the control parameters (e.g. the reorder point \(s\), the batch size \(Q\)). Some papers [52,55] also indicate how these (or some of these) control parameters should be chosen, such that an additional constraint is satisfied. E.g. Sherbrooke [55] gives an optimization procedure to allocate the safety stock among the facilities given any total system stock investment. Schneider, Rinks and Kelle [52] derive a heuristic that aims at minimizing the holding costs under a service level constraint.
Below we use the following additional notation:

\[ L_i := \text{the effective lead time of retailer } i, \]
\[ \lambda_i := \text{rate of customer arrivals at retailer } i, \]
\[ \mu_i := \text{mean lead time demand at retailer } i, \]
\[ \sigma^2_i := \text{variance of lead time demand at retailer } i, \]
\[ B_i := \text{the expected number of products back-ordered at stockpoint } i, \]
\[ O_i := \text{number of outstanding orders at stockpoint } i, \]

where stockpoint \( i \) refers to retailer \( i \) for \( 1 \leq i \leq M \), and refers to the depot for \( i = 0 \). In Section 3.1 we address systems under continuous review, and in Section 3.2 we address the systems under periodic review.

### 3.1. Continuous review policies

In most papers concerning continuous review policies demand is assumed to be a (compound) Poisson process with rate \( \lambda_i \). The inventory at stockpoint \( i \) is controlled by an installation stock \((s_i, nO_i)\) policy, where not only retailers back-order excess demand, but also the depot when it is not able to fill all retailers.

One of the first papers which models the system interactions in divergent multi-echelon environments is the METRIC model of Sherbrooke [55]. METRIC stands for "Multi-Echelon Technique for Recoverable Item Control", and analyzes how to maintain an inventory of repairable items (items usually with high cost and low demands) using an \((S_i - 1, S_i)\) policy, which for the case of unit demand corresponds to an order-point, order-quantity policy with \( Q_i := 1 \) and \( s_i := S_i - 1 \).

The work on \((S_i - 1, S_i)\)-systems like METRIC can, in general, be viewed in either installation or echelon stock terms. However, in Section 3.1.1 we elaborate on the METRIC model (for repairable items), since this model constitutes the basis for a lot of installation stock models which analyze a model similar to METRIC, but for consumable items. Since these consumable items usually have low cost and high demand, the order size required for these items generally will exceed 1. In those cases an order-point, order-up-to policy cannot be described by an equivalent order-point, order-quantity policy. In Section 3.1.2 we address these installation stock models for consumable items. In Section 3.1.3 we address specific modeling problems concerning the analysis of these models.

#### 3.1.1. Repairable items

METRIC models a supply system consisting of a repair depot and an arbitrary number of operating bases. The depot and the bases maintain an inventory of spare parts. The demand for spare parts is only generated at the bases and is assumed to be compound Poisson. For ease of presentation we suppose that at base \( i \) failures occur according to a Poisson process with rate \( \lambda_i \). When an item fails at base \( i \), with some probability \( r_i \) it can be repaired at that base according to an arbitrary probability distribution of the repair time, otherwise it must be returned to the depot, where it is repaired according to some other arbitrary repair time distribution. It is assumed that there every item can be repaired and that lateral supply between bases is not possible. So, whenever a failure occurs the base supplies, if possible, an item from the stock on-hand. Immediately after such a failure this item is sent to the repair unit of the base, or sent to the depot for repair. When a failed item is sent to the depot, at the same time the base places a resupply request on the depot. As a consequence, the items are not batched for repair and resupply requests.

The main METRIC goal is to allocate the available safety stock over the bases and depot such that the costs induced by the back orders are acceptable, i.e., equals a prespecified amount. Hence, we need to compute the expected number of products back-ordered at a base, say base \( i \), and the depot. This number is denoted by \( B_i \) and \( B_0 \), respectively. Denote the number of outstanding orders at base \( i \) and at the depot by \( O_i \) and \( O_0 \), respectively, it follows that

\[ B_i = E\{\max[0, O_i - S_i]\}, \quad i = 0, 1, \ldots \quad (1) \]

When the distribution of \( O_i \) is known \( B_i \) can easily be computed from (1). We also need the distribution of the outstanding orders at a base \( i \) and at the depot. Since a failed item can be repaired both at the base and at the depot, an outstanding order at a base can be an item on order at this base or an item on order at the depot. The outstanding orders at each base can simply be analyzed by a single-echelon model. The number of outstanding orders of each base at the depot and the total number of outstanding orders at the depot can be analyzed as a divergent two-echelon system as introduced in Section 2.1. We shall address these two analyses separately.
The depot analysis

An operating base is modeled as a retailer. A demand occurrence for a single item at a retailer can be seen as the occurrence of a failure at a base, which cannot be repaired at the base but is sent to the depot for repair. The requests from all bases correspond to the demand at the depot. If on-hand stock at the depot is sufficient, a spare item is sent to the base after \( L_i \) time units. Besides sending a resupply request, the failed item is sent to the repair unit for repair. The base not only plays the role of a retailer demanding spare items, but also of a supplier supplying failed items to the depot. The in-transit time of a failed item from base \( i \) to the depot, plus the repair time at the depot equals the lead time \( L_0 \) and is independent of base \( i \). Since with probability \( 1 - \tau_i \) a failed item will be repaired at the depot, the demand process of retailer \( i \) is a Poisson process with rate \( (1 - \tau_i) \lambda_i \). Hence, the demand process at the depot is a Poisson process with rate \( A := \sum_i (1 - \tau_i) \lambda_i \). Since the depot uses a one-for-one replenishment policy, the demand process at the supplier is identical to the demand process at the depot. We assume ample repair capacity, i.e., immediately after the arrival of a failed item at the depot the repair starts. Hence, the amount of outstanding orders at the depot \( O_0 \) is identical to the occupancy level, i.e., the number of busy servers, in an M/GI/\( \infty \) queue. According to the theorem of Palm [47], the steady-state probability distribution of the occupancy level in an M/GI/\( \infty \) queue is Poisson with rate \( A E Lo \), if the arrival rate equals \( A \) and the mean service time equals \( E Lo \). In [29] Palm’s theorem is extended to the compound Poisson demand case, under the condition that the repair times of all items in one demand batch are identical.

The base analysis

Consider the repair unit at base \( i \). The arrival rate of failed items equals \( \tau_i \lambda_i \). Assuming ample repair capacity and that the repair times are i.i.d. following an arbitrary distribution with mean \( T_i \), Hence, from Palm’s theorem, the number of outstanding orders at the repair unit of base \( i \) is Poisson distributed with mean \( \tau_i \lambda_i T_i \). However, some of the failed items are sent to the depot for repair. The arrival rate of resupply requests of base \( i \) at the depot equals \( (1 - \tau_i) \lambda_i \). The effective lead time \( L_{i,k} \) of the \( k \)th order of base \( i \) is at least the shipping time \( L_i \), but possibly an additional waiting time \( W_{i,k} \) (due to material shortage at the depot),

\[
L_{i,k} = L_i + W_{i,k}, \quad i = 1, 2, \ldots, M, \quad k = 1, 2, \ldots, \tag{2}
\]

where \( W_{i,k} \) denotes the additional waiting time of the \( k \)th order of base \( i \). When both the stock-outs at the depot are filled on a first-come-first-served (FCFS) basis, and the demand at base \( i \) is Poisson distributed, \( W_{i,k} \) is independent of the base placing the order.

For the special case where the depot lead time equals a constant \( L_0 \), Sherbrooke [57] derived the cumulative distribution function \( F \) of \( W_{i,k} \) (denoted by \( W \)). For \( S_0 > 0 \),

\[
F_W(w) = \sum_{k=0}^{S_0-1} \frac{(-A(\lambda_0 - w))^k}{k!} e^{-A(\lambda_0 - w)}, \quad 0 \leq w \leq \lambda_0
\]

and for \( S_0 = 0 \), \( F_W(w) = 0 \) for \( 0 \leq w < \lambda_0 \) and \( F_W(\lambda_0) = 1 \). This result can be verified immediately from the observation of Asmussen [2]: a product ordered by the depot is used to fill the \( S_0 \)th demand following this order. Hence the cumulative distribution function \( F \) of the time elapsed between the placement of an order and the occurrence of the \( S_0 \)th demand following this order corresponds with an Erlang \((A, S_0)\) distribution.

We now return to the relation stated in (2). We like to emphasize that the successive waiting times, \( W_{i,k} \) and \( W_{i,k+1} \), are identically distributed, but are also correlated. Hence, we cannot use Palm’s theorem to obtain the distribution of the number of outstanding orders of base \( i \) at the depot. Now METRIC makes the following approximation: disregard the correlation between successive waiting times by defining

\[
L_{i,k} := L_i + \mathbb{E} W_{i,k}, \quad i = 1, 2, \ldots, M, \quad k = 1, 2, \ldots, \tag{3}
\]

Under this assumption Palm’s theorem can be used. The number of outstanding orders of base \( i \) at the depot is Poisson distributed with mean \( \lambda_i \mathbb{E} L_i \). Hence, \( O_i \) is Poisson distributed with mean \( \lambda_i [\tau_i T_i + (1 - \tau_i) \mathbb{E} L_i] \). The only aspect which needs to be analyzed is \( \mathbb{E} W_{i,k} \).
As already mentioned, $O_0$ is Poisson distributed with mean $\lambda \in L_0$. From (1) we obtain $B_0$. Next, by applying Little's well-known formula [43] we have

$$\mu_w = \frac{B_0}{\lambda},$$

(4)

where $\mu_w$ is the expected waiting time of a product ordered at the depot. Since the demand at a base follows a Poisson process, the waiting time $W_{i,k}$ is independent of the base $i$ (and also of $k$). Hence, using the Poisson Arrivals See Time Averages (PASTA) property of Wolff [74], $E W_{i,k} = \mu_w$. The analysis becomes more complicated when the demand process at each base is a compound Poisson process, since then the waiting time $W_{i,k}$ also depends on the base placing the order and Little's formula does not hold anymore (except for some specific cases). Further on the applicability of Little's formula will be discussed.

This concludes our analysis of METRIC. For a more extensive analysis we refer to Sherbrooke [59]. He presents METRIC and its assumptions extensively, although, in our opinion, he disregards to discuss the implicit assumption that the rate $\lambda_i$ is independent of the number of failures at base $i$. Indeed, when the number of items at base $i$ is large the decrease of $\lambda_i$ will be negligible, otherwise the impact of this assumption is a priori not clear.

Extensions

Several extensions of METRIC have been developed over the years. For a more extensive review we refer to [3,13,44,46]. We shall briefly discuss some of these extensions.

Graves [30] and Slay [62] developed the so-called VARI-METRIC model which uses another approximation to determine $O_i$. An important difference with METRIC is that the order-and-shipment times $L_i$ are assumed to be deterministic. Using a result in Simon [61] (he determines the distribution of $O_i$ when the back orders at the depot are filled on a FCFS basis), they compute the first two moments of $O_i$. Graves [30] and Slay [62] propose to fit a negative binomial distribution on these moments, to approximate the distribution of $O_i$. Graves [30] compared the performance of this approximation and the METRIC approximation with exact results obtained by computing the required stockage levels for four bases such that every base meets his predefined service level $\alpha$, i.e., $Pr\{O_i < S_i\} \geq \alpha$. It appears that the METRIC approximation computes too low stockage levels in 11.5% of the 2304 cases considered, while the negative binomial approximation results in wrong stockage levels in only 0.9% of these cases.

An essential extension to METRIC concerns the incorporation of the multi-indenture relationship between end items and their comprising modules. Consider an aircraft engine consisting of a number of replaceable modules. METRIC minimizes the expected back orders of all items (both engines and modules), while in practice only shortages of end items (engines) affect the downtime of the system. Sherbrooke [56] was the first to recognize this multi-indenture relationship. Muckstadt [45] extended the METRIC to a multi-indenture model also called MOD-METRIC. Sherbrooke [58] extends the VARI-METRIC model by taking the multi-indenture relationship into account. In Lee [41] and Axsäter [1] the basic model is extended by allowing lateral transshipments (between the bases).

The METRIC based models discussed so far focus on characterizing the steady-state behavior of the inventory levels for a given ordering policy, using the steady-state distribution (or an approximation thereof) to determine the average costs associated with the policy. For Poisson demand and deterministic lead times Axsäter [2] provides a more efficient and direct method to find the optimal inventory policy minimizing an inventory cost function that reflects costs incurred on an average unit. This approach does not require the METRIC assumption, neither does it provide any information on the steady-state distribution of inventory levels, necessary to determine service levels. Recently Axsäter, Forsberg and Zhang [4] proposed an alternative approach to determine "appropriate" order-up-to levels $S_i$ in case of compound Poisson demand and deterministic lead times $L_i$. They replace the compound Poisson demand process at every retailer by an "equivalent" Poisson demand process such that the ratio between the mean and standard deviation is the same as for the real distribution. As a consequence the demand process at the depot also becomes a Poisson process. Next, the algorithm of Axsäter [2] provides the optimal order-up-to levels in the adapted model (Poisson demand), which are used to compute the order-up-to levels in the original model (i.e., compound Poisson demand). This ap-
proach can easily be adapted when every stockpoint is controlled by an order-point, order-quantity policy (see Section 3.1.2). Also, extension to systems with three or more echelons is straightforward, yet no numerical results are available.

3.1.2. Consumable items

The METRIC model cannot be extended to the case where the batch sizes exceeds 1, since Palm's theorem requires \( Q = 1 \). For items with a high demand it makes sense to have a batch size \( Q \) larger than 1 due to the ordering costs which have to be paid for every order. Deuermeyer and Schwarz [21] extend the decomposition method of METRIC to analyze the divergent two-echelon \((s, nQ)\) system for consumable items. They consider the case of Poisson demand at retailer \( i \) with rate \( \lambda_i \).

Suppose retailer \( i \) places an order of \( Q_i \) at the depot. This order arrives after an effective lead time \( L_i \) given by (2). In contrast with METRIC, which does not require a deterministic \( L_i \), in the model of Deuermeyer and Schwarz [21] deterministic lead times \( L_i \) are assumed. Lot-splitting at the depot is prohibited. Back orders at the depot are filled on a FCFS basis.

The analysis is based on the METRIC approximation, i.e., the effective lead times are defined by (3). It is important to realize that this enables to decompose the divergent two-echelon system into several single location inventory systems. Before we elaborate on the analysis, we show how to determine the service measure \( \gamma_i \) and \( \gamma_i' \), respectively. Deuermeyer and Schwarz [21] give a computationally convenient normal approximation of the expressions derived in Hadley and Whitin [31]:

\[
\beta_i = 1 - \frac{\alpha(s_i^+) - \alpha(s_i + Q_i) - (-s_i)^+}{Q_i},
\]

\[
\gamma_i' = 1 - \frac{(s_i^+) - \beta(s_i + Q_i) - (-s_i)^+(\frac{1}{2}(s_i + 1) - \alpha(0))}{\lambda_i Q_i},
\]

where

\[
\alpha(v) := \sigma_i \Phi\left(\frac{v - \mu_i}{\sigma_i}\right) - (v - \mu_i) \Phi\left(\frac{v - \mu_i}{\sigma_i}\right),
\]

\[
\beta(v) := \frac{1}{2} \left[ \sigma_i^2 \Phi\left(\frac{v - \mu_i}{\sigma_i}\right) - (v - \mu_i) \Phi\left(\frac{v - \mu_i}{\sigma_i}\right) \right],
\]

with \( x^+ = \max(0, x) \), \( \sigma_i := \sqrt{\mu_i} \) and \( \mu_i := \lambda_i \mathbb{E} L_i \). \( \Phi(\cdot) \) and \( \Phi(\cdot) \) are the standard normal density and complementary distribution, respectively. The service measure \( \gamma_i' \) is the modified fill rate definition introduced in [51], which is defined as one minus the amount of cumulative back orders per unit time divided by the mean demand per unit time. Notice the difference with our \( \gamma \)-definition in Section 2.4 which is based on the behavior of the stock at the end of a replenishment cycle, instead of the behavior per time unit.

The depot analysis

The analysis of the depot is hard due to the complexity of the demand process. In general this demand process is a non-stationary compound point process. However, Deuermeyer and Schwarz [21] only consider the case where all order sizes \( Q_i (i = 1, \ldots, M) \) are identical, say \( Q_r \). Hence, the demand process at the depot becomes a counting process. Furthermore, the lead time \( L_0 \) is deterministic, and the order size \( Q_0 \) is a multiple of \( Q_r \). They approximate the demand process at the depot as the order process of one artificial retailer, with a demand process which is the superposition of the \( M \) Poisson processes at the retailers, a Poisson process in itself with rate \( \Lambda := \sum_{i=1}^{M} \lambda_i \).

To illustrate the difference between the real demand process at the depot and the demand process resulting from this approximation we consider a divergent two-echelon model with two end-stockpoints. Fig. 2 depicts the demand process of both retailers. A circle represents the arrival of a customer, and a filled circle means that an order is placed at the depot. Also assume that \( Q_r = 2 \), hence for each retailer an order is placed at the depot after every two customer-arrivals at this retailer. The demand process at the depot follows by superposing the order-processes at every retailer. The artificial retailer places an order after every two customer-arrivals (irrespective of where the customers arrive). From Fig. 2 it becomes clear that the

![Fig. 2. The demand process at retailer 1, retailer 2, the depot and the depot (using the approximation), respectively (\( M = 2 \) and \( Q_r = 2 \)).](image-url)
actual demand process at the depot and the order process of the artificial retailer are different.

Using well-known approximations from renewal theory we obtain the mean and variance of the depot lead time demand, denoted by $\mu_0$ and $\sigma_0^2$, respectively.

$$\mu_0 = \frac{A_0}{Q_r} + \frac{1 - Q_r}{2Q_r} \quad \text{and} \quad \sigma_0^2 = \frac{A_0}{Q_r^2}. \quad (7)$$

Finally, Deuermeyer and Schwarz [21] fitted a normal distribution on $\mu_0$ and $\sigma_0^2$.

The retailer analysis

Under the METRIC approximation every retailer can be modeled as a single location under $(s_i, nQ_i)$ control with a fixed lead time $E L_i$. In order to determine the effective lead time $E L_i$, $E W$ is computed as in Section 3.1.1. This means that we first determine $B_0$ by substituting (7) in (5), and next use Little’s formula to obtain $E W = B_0/\Lambda$.

Extensions

Svoronos and Zipkin [64] followed the same approach as Deuermeyer and Schwarz [21]. The key innovation of their paper is not to treat the effective lead time of a retailer as deterministic, but approximate both the mean and the variance of this effective lead time, and fit a negative binomial distribution on these two moments. With the result of Svoronos [63], who extended the results of Hadley and Whitin [31] by considering a stochastic leadtime, an expression for $B_i$ can be obtained. Furthermore, Svoronos and Zipkin [64] derived the exact mean and variance of the depot lead time demand, and fitted the better performing Mixed Translated Poisson (MTP) distribution (i.e., with probability $1 - p$ we obtain a shifted Poisson distribution with mean $v$ and shift $A$, and with probability $p$ we obtain a shifted Poisson distribution with mean $v$ and shift $A + 1$) on these moments. In [42] a two moment approximation is used to fit the depot lead time demand. However, to analyze the retailers they use a quite different approach: instead of modeling the lead times as an intermediate step, they estimate the back orders at the retailers directly from the back orders at the warehouse, using a disaggregation procedure similar to that of Graves [30]. A numerical study performed by Svoronos and Zipkin [64] indicates that their model is more accurate than those of Deuermeyer and Schwarz [21] and Lee and Moinzadeh [42].

New developments in the analysis of these depot-retailer systems have appeared recently. In [11] every stockpoint uses an $(s, S)$ policy. The demand process at each retailer follows a stationary stuttering Poisson process (i.e., a compound Poisson process with geometrically distributed order-sizes) and all retailers are identical. Their approach to estimate the expected fill rate fundamentally differs from (5), since the expected back orders at the end of each reordering cycle at retailer $i$ is determined by conditioning on the event of a depot stock-out, namely

$$(1 - p)B_1 + pB_2,$$

where $p$ is the steady-state probability that the depot is unable to satisfy an order of retailer $i$, and $B_1$ and $B_2$ are the expected back orders per cycle at retailer $i$ given that the depot is unable/able to satisfy the order, respectively. A more general model is studied in [12], where stockpoint $i$ uses an $(s_i, S_i)$ policy. The demand at every retailer is a stationary Poisson process. They provide an expected upper bound for the stock-outs at the depot per unit time, which is used to approximate the reorder point and the order-up-to level at the depot, in order to minimize average depot cost.

3.1.3. Specific modeling problems

"When a retailer uses an $(s_i, nQ_i)$ replenishment policy or the demand processes at these retailers are compound Poisson the analysis of the model becomes far more complex. This is due to two phenomena. First of all, the analysis above relies on the applicability of Little’s formula, which does not hold in general. Secondly, the assumptions under which models are analyzed so far are such that replenishment orders are satisfied completely or back-ordered completely. In general this does not hold. Let us look into these phenomena in more detail."

The applicability of Little’s formula

As shown in the retailer analysis in Section 3.1.2, an expression for $E W$ is derived from Little’s formula. It turns out that the applicability of Little’s formula is restricted to particular cases. The following are the most studied cases in the literature where the formula can be applied:

(i) Unit demand (e.g. METRIC and its extensions).
(ii) Deterministic demand sizes \( D \) and reorder level \( s \) is a multiple of \( D \) (e.g. Svoronos and Zipkin [64] assume identical order batches at the retailers, resulting into deterministic demand sizes at the depot).

(iii) Exponential demand sizes.

With compound Poisson demand Little’s formula cannot be generally applied to derive the average waiting time \( \mu_w \). This is caused by the fact that the sizes of subsequent back orders are not i.i.d. with the same distribution as the demand distribution \( F_D \). More precisely, the first back order that occurs at a stock-out occasion has a different distribution, whereas all subsequent back orders until the next replenishment are i.i.d. with distribution \( F_D \). In case some demand can still not be filled by this replenishment a similar situation occurs with the customer demand size that is only partially filled as opposed to the demands that are not filled at all. We can see that for the above three cases all back orders are i.i.d. with distribution \( F_D \). For exponential demands this is caused by the lack of memory of the exponential distribution.

To appreciate the errors caused by applying Little’s formula we present some simulation results for a single location \((s, nQ)\) model with compound Poisson demand with rate \( \lambda \) and a customer batch size of \( D \). This rate \( \lambda \) equals 1 and batch size \( D \) has a mean of 100 and a squared coefficient of variation \( c^2_D \). We varied \( c^2_D \) as 0.5, 1, 1.5 and 2. The reorder level \( s \) is determined for a fill rate \( \beta \), where \( \beta \) is varied as 0.50, 0.75, 0.9. The order size \( Q \) equals 1000. Order-splitting is allowed. In Table 1 we compare \( \mu_w \) with \( \mu_w^{\text{Little}} \), where

\[
\mu_w^{\text{Little}} := \frac{B}{\lambda E D},
\]

where \( B \) is the expected number of backlogged products. Furthermore we provide an approximation for \( \mu_w \) based on the assumption that the first back order has the same distribution as the stationary residual life distribution associated with \( F_D \). This yields the following approximation for \( \mu_w \),

\[
\mu_w^{\text{Approx}} := \frac{B}{\lambda E D} + \pi \left( \frac{E D^2}{2 E D} - E D \right), \tag{8}
\]

where \( \pi \) equals the probability that the net stock (physical stock minus back orders) is negative. In [17] an accurate approximation for \( \pi \) is derived.

Table 1

<table>
<thead>
<tr>
<th>( \beta )</th>
<th>( c^2_D )</th>
<th>( \mu_w )</th>
<th>( \mu_w^{\text{Little}} )</th>
<th>( \mu_w^{\text{Approx}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1.42</td>
<td>1.30</td>
<td>1.42</td>
</tr>
<tr>
<td>1.0</td>
<td>1.37</td>
<td>1.38</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.32</td>
<td>1.44</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.24</td>
<td>1.44</td>
<td>1.26</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.5</td>
<td>0.51</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>1.5</td>
<td>0.48</td>
<td>0.53</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.45</td>
<td>0.54</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
<td>0.16</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>1.0</td>
<td>0.15</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>1.5</td>
<td>0.15</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>2.0</td>
<td>0.15</td>
<td>0.18</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

From the results of Table 1 we see that this adaptation of Little’s formula yields a robust and accurate approximation. Also the simulation reveals that \( \mu_w \) is rather insensitive with regard to \( c^2_D \). More research needs to be done to investigate this phenomenon.

**Heterogeneous systems and lot splitting**

In most papers on two-echelon inventory systems under installation stock policies with lot sizing the following assumptions are made:
- Lot sizes are identical for all retailers.
- In case of shortages at the depot the retailer order is delayed until it can be satisfied completely.

The latter assumption is referred to as the no-lot-splitting assumption. The impact of these assumptions has never been explicitly investigated. In practice the assumption of identical lot sizes of retailers (or customers) is invalid. E.g. suppose that the retailers are in fact wholesalers, power retailers and other stock points of the company that owns the depot. Then the EOQ formula tells us that lot sizes depend on the demand at each customer base and the cost structure of each base. These are in practice quite different.

It is quite common to split lots in case of material shortages. Lot-splitting also impacts the cash flow of the supplier companies and their capital tied up in inventory. Indeed, if lot-splitting is applied remnant stock does not occur in case of shortages and customers can be invoiced immediately for the products shipped. Moreover, customers may be able to start op-
operations before the complete order arrives. To illustrate the importance of further research with respect to these common assumptions we give an example of a heterogeneous system with different lot sizes for the retailers. We compare the situations with and without lot splitting. In Table 2 we show the model parameters of the system considered and the results obtained by simulation. These reveal that the lot-splitting assumption can have a large impact on the attained fill rate. As already mentioned, due to heterogeneity the waiting time of an order at the depot depends on the retailer placing the order (see Table 2). A topic of further research may concern when lot-splitting will have a considerable impact on the performance of the supply chain.

### Table 2

<table>
<thead>
<tr>
<th>$i$</th>
<th>$s_i$</th>
<th>$Q_i$</th>
<th>$\beta_i^{NLS}$</th>
<th>$\beta_i^{LS}$</th>
<th>$\mu_i^{NLS}$</th>
<th>$\mu_i^{LS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
<td>0.25</td>
<td>0.917</td>
<td>0.888</td>
<td>0.901</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100</td>
<td>2.25</td>
<td>0.888</td>
<td>0.901</td>
<td>0.438</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100</td>
<td>0.06</td>
<td>0.857</td>
<td>0.990</td>
<td>0.478</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>25</td>
<td>4.00</td>
<td>0.895</td>
<td>0.990</td>
<td>0.478</td>
</tr>
</tbody>
</table>

Various assumptions are needed to analyze this interaction making the analysis system specific.

Van der Heijden [32] considers a model where the inventory at stockpoint $i$ is controlled by an installation stock ($R_i, S_i$) policy allowing for a different review period at each retailer. Hence, the model in fact incorporates lot sizing. It is assumed that the review period at the depot is an integer multiple of the review periods at the retailers, since otherwise the demand process at the depot results in a non-stationary demand process (compare the case of non-identical $Q_i$ in Section 3.1.2). In practice this assumption is usually no restriction, since the inventory is inspected more frequently at stockpoints downstream rather than upstream. An important feature of this model is that replenishment orders may be issued on different points in time. For example, one retailer reviews its inventory each Monday, the other one each Friday. We note that the approach of Van der Heijden [32] is less restrictive than the previous ones [21,64], allowing for compound Poisson demand and different retailer control strategies ($R_i$ and $S_i$).

Like in most installation stock models a key part in the analysis is the determination of the waiting time. The first two moments of $W_{ik}$ are approximated by using the results of Van der Heijden and De Kok [69] and the PDF method of De Kok [17]. Unlike METRIC and Deuermeyer and Schwarz [21], these two moments for every retailer $i$ are determined separately (so the waiting time distribution depends on $i$). Next, the effective lead time distribution is approximated by a Coxian distribution using a two-moment fit. The PDF method is used to determine the fill rate and the mean physical stock at a stockpoint $i$. Simulation results indicate that the approximation is quite accurate for the most important performance measures (external fill rates and mean physical stock in the network). Errors on other performance measures (internal fill rates and
the mean physical stock in a specific stockpoint) can be larger.

Schneider, Rinks and Kelle [52] consider a similar inventory system as Van der Heijden [32]. In their case the inventory at stockpoint \( i \) is controlled by an installation stock \((R, s_i, S_i)\) policy. To analyze this two-echelon inventory network Schneider, Rinks and Kelle [52] use a similar approach as discussed in Section 3.1.1, to decompose the whole system into several single location systems. The aim is to determine the control parameters such that the long-run average costs are minimized subject to the condition that every retailer attains a predetermined fill rate. The service level under consideration is the fill rate. We shall give a brief overview of their approach.

Due to the decomposition the depot is treated as a single location system. In order to model the long-run average costs the lead time demand distribution is needed. This distribution is approximated by fitting a negative binomial distribution on the first two moments. These first two moments are derived in [53]. Since the demands at the depot are auto-correlated the second moment is rather cumbersome to obtain. Hence, Schneider, Rinks and Kelle [52] use an approximate expression derived by Ehrhardt, Schultz and Wagner [25].

Next Schneider, Rinks and Kelle [52] derive an approximation for the lead time demand distribution of every retailer, and then give an expression for the long-run average cost of every retailer. They approximate the two moments of the retailers effective lead time assuming that in case of a depot stock-out all orders are delayed. In reality this may not be the case. There may be enough stock to satisfy some orders from retailers, but not others. Under this condition

\[
\Pr\{L_i = l_i + j\} = a_j - a_{j-1},
\]

\( j = 0, 1, \ldots, l_0 + 1, \) (9)

where \( a_j (i = 1, \ldots, l_0) \) is the probability that a depot stock-out lasts at most \( i \) review periods. Furthermore, \( a_{-1} := 0 \) and \( a_{l_0+1} := 1 \). Schneider [50] provides an approximation for \( a_i \) for large \( S_0 - s_0 \). From (9) they approximate the first two moments of \( L_i \). A negative binomial distribution is fitted on the lead time demand distribution of every retailer \( i \).

The trade-off between the safety stock at the depot and the safety stock at the retailers is modeled by imposing a given \( \alpha \) service at the depot. When \( \alpha \) is small, this causes a delay of the retail-orders. In order for the retailers to meet their target service levels they are forced to increase their safety stock. As a consequence, the expected total stocking costs will increase. On the other hand, when \( \alpha \) is large, the depot holds a lot of stock on-hand, which increases holding costs. The optimal policy can be obtained by the Lagrangian method. Approximation techniques, such as the power approximations of Ehrhardt [24], are used to obtain the control parameters of this policy.

4. Echelon stock policies

A central role in the analysis of multi-echelon inventory systems is the concept of echelon stock, originally introduced by Clark [14] and later used by Clark and Scarf [15] to establish the optimality of echelon order-up-to policies in a serial system. In order to prove the optimality of these policies we need that there are no setup costs at all but the most upstream stockpoint. A recent paper of Chen and Zheng [10] also studies this serial system, however, setup costs are incurred at all stockpoints. Therefore they suggest to control every stockpoint by an echelon stock \((s, nQ)\) policy. A recursive procedure is developed to compute the steady-state echelon inventory positions, which can be used to evaluate the long-run average holding and back-order costs as well as service measures. Their results apply to both the continuous review systems with compound Poisson demand and periodic review systems with independent, identically distributed demands. Besides their paper and the paper by Van Donselaar [66] to our knowledge little results are available on continuous review echelon stock policies for multi-echelon systems. Hence, in this section we concentrate on the problems typical for periodic review echelon stock policies. First of all we need to have some rationing policy which allocates the available echelon stock at the depot (i.e., the echelon inventory position minus stock in transit to depot) over the retailers. In Section 4.1 we shall discuss several rationing policies. We emphasize that these rationing policies do not ration the stock on hand at the depot but the echelon stock at the depot. Then it may occur that the echelon inventory position of a retailer just after rationing is less than just before rationing. In Sec-
tion 4.2 we elaborate on this phenomenon, which is referred to as imbalance.

4.1. Rationing policies

In discussing rationing policies, we focus on practically useful policies which enable to keep track of material flows throughout the network, which allows us to derive expressions for the system performance. The duration of the review period at the depot and at the retailers are equal, and the review moments are synchronized with the arrival of replenishments from the supplying stockpoint. If we also assume constant lead times without loss of generality a review period corresponds with one period. In Section 4.1.1 we elaborate on the well-known Fair Share (FS) rationing policy. In Section 4.1.2 we discuss the more general Appropriate Share (AS) rationing policy. We also present a numerical study, which yields insight into the performance of AS rationing and the impact of the imbalance on the performance. In Section 4.1.3 we discuss the Priority rationing policy of Lagodimos [38]. In Section 4.1.4 some extensions and other rationing policies are discussed. Section 4.1.5 yields expressions to actually compute the service levels introduced in Section 2.4 for any rationing policy. In this section we introduce some additional notation:

\[ \mu_i := \text{mean demand at retailer } i \text{ during one period}, \]
\[ \sigma_i^2 := \text{variance demand at retailer } i \text{ during one period}, \]
\[ U_t^i := \text{the projected net inventory of retailer } i \text{ at time } t + l_i + 1, \]
\[ U_t := \text{the system-wide projected net inventory; } U_t = \sum_{n=1}^{M} U_t^n, \]
\[ p_i := \text{allocation fraction of retailer } i \text{ of rationing at the depot}, \]
\[ \Delta := S_0 - \sum_{n=1}^{M} S_n. \]

4.1.1. Fair share rationing

Suppose at the end of an arbitrary period, say at time \( t - l_0 \), the depot places an order at the supplier to raise its echelon inventory position to \( S_0 \). Then this order arrives at time \( t \) (since the depot lead time equals \( l_0 \) periods). Just after this arrival (but prior to material rationing), the inventory position of all retailers plus the on-hand stock at the depot equals \( S_0 - D_{t-l_0,t} \). Hence, in order for the retailers to raise their inventory position to their order-up-to levels, \( S_0 - D_{t-l_0,t} \) has to exceed \( \sum_{i=1}^{M} S_i \):

\[ \Delta \geq D_{t-l_0,t}. \]

Notice that \( \Delta \) influences the depot operation. If \( \Delta \leq 0 \) the depot will not hold any stock. This implies that when a product arrives at the depot it is immediately allocated to the retailers. If \( \Delta = \infty \), the system decomposes into \( M \) single location systems working in parallel.

For the determination of the control parameters usually one may seek to achieve some pre-determined target service levels at the retailers. Otherwise, these are determined by minimizing a cost-criterion (see [70] for a recent review).

An important contribution in the development of reasonable rationing policies is given by Eppen and Schrage [26]. This paper examines the inventory system as depicted in Fig. 1. Each retailer demand is normally distributed, stationary and independent from that of other retailers. The lead times between the depot and the retailers are fixed and identical and every retailer has the same target stock-out probability \( \alpha \). The depot is not allowed to hold any stock, but merely serves as a coordinator, i.e., \( \Delta := 0 \). Therefore at every review moment rationing is required. Eppen and Schrage [26] introduced the well-known Fair Share (FS) rationing policy, which was later extended by Van Donselaar and Wijngaard [67] and Lagodimos [38]. The FS rationing policy rations the available material so as to maintain all the end-stockpoints at a balanced position: all end-stockpoints have the same non-stock-out probability \( \alpha \). When the period demands of every retailer \( i \) are normal random variables \( N(\mu_i, \sigma_i^2) \) (uncorrelated in time), then (cf. Lagodimos [38])

\[ I_t^i - (l_i + 1) \mu_i \sigma_i \sqrt{l_i + 1} = \sum_{n=1}^{M} \left[ I_t^n - (l_n + 1) \mu_n \right] \sigma_n \sqrt{l_n + 1}. \]

So, using this expression and the definition of \( U_t \) yields

\[ I_t^i = (l_i + 1) \mu_i + \frac{\sigma_i \sqrt{l_i + 1}}{\sum_{n=1}^{M} \sigma_n \sqrt{l_n + 1}} U_t. \] (10)

4.1.2. Appropriate share rationing

De Kok [18] extends the analysis of Eppen and Schrage [26] to arbitrary demand functions, arbitrary
service criteria, and non-stationary demand. Furthermore, the lead times between the depot and the retailers, as well as the target service levels may differ. Unlike Eppen and Schrage [26], the approach of De Kok [18] is based on the use of service criteria instead of cost criteria. However, the depot is still not allowed to hold any stock. This restriction is relaxed in a later paper of De Kok, Lagodimos and Seidel [20], which introduces the Appropriate Share (AS) rationing policy. The purpose of AS rationing is to ensure that a prespecified target service level can be attained at a retailer.

In order to explain AS rationing policy properly we introduce $U^i_t$, which will be referred to as the projected net inventory of retailer $i$,

$$U^i_t := I^i_t - (l_i + 1) \mu_i. \quad (11)$$

$U^i_t$ represents an estimate (made at time $t$) of the net inventory (stock on hand minus back orders) at retailer $i$ at time $t + l_i + 1$, given the inventory position of retailer $i$ at time $t$ just after rationing. The system-wide projected net inventory $U_t$ (at retailer level) is defined by $\sum_{i=1}^{M} U^i_t$.

At the beginning of a review period $t$ there are two possibilities:

(i) $t \in T_S := \{ t \mid \Delta \geq D_{t-l_0,t} \}$. All the retailers are able to raise their inventory positions to their order-up-to levels. Thus, $\sum_{i=1}^{M} S_i$ is allocated to the retailers, and the remainder $\Delta - D_{t-l_0,t}$ is retained at the depot. Hence, $U_t$ equals the planned cumulative safety stock of the retailers,

$$U_t = \sum_{n=1}^{M} (S_n - (l_n + 1) \mu_n), \quad t \in T_S. \quad (12a)$$

(ii) $t \in T_R := \{ t \mid \Delta < D_{t-l_0,t} \}$. All depot inventory is allocated to the retailers and rationing is required. Since the “shortage” equals $D_{t-l_0,t} - \Delta$, we have

$$U_t = \sum_{n=1}^{M} (S_n - (l_n + 1) \mu_n) - (D_{t-l_0,t} - \Delta), \quad t \in T_R. \quad (12b)$$

If $t \in T_R$ we have to decide how to ration the available stock over the retailers. AS rations this available depot inventory so that the projected net inventory $U^i_t$ over the system-wide projected net inventory $U_t$ equals a prespecified fixed allocation fraction $p_i$,

$$p_i := \frac{U^i_t}{U_t}, \quad t \in T_R. \quad (13)$$

Clearly, we need that $\sum_{i=1}^{M} p_i = 1$. Rewriting (13) with (11) and (12b) yields

$$p_i = \frac{I^i_t - (l_i + 1) \mu_i}{\sum_{n=1}^{M} (S_n - (l_n + 1) \mu_n) - (D_{t-l_0,t} - \Delta)}, \quad t \in T_R.$$  

Substitution of (11) in (13), and some elementary algebra yields

$$I^i_t = \begin{cases} (l_i + 1) \mu_i + p_i U_t, & t \in T_R, \\ S_i, & t \in T_S. \end{cases} \quad (14)$$

This expression can be interpreted as follows. When the depot inventory is sufficient every retailer raises his inventory position to the order-up-to level. Otherwise, the inventory position of retailer $i$ equals the expected demand at retailer $i$ during the replenishment lead time plus a review period, plus a fraction $p_i$ of the system-wide projected net inventory.

With AS rationing the allocation fractions $p_i$ still can be chosen freely. In a restricted version of AS rationing, referred to as Consistent Appropriate Share (CAS) rationing [20], $p_i$ is chosen so that

$$p_i := \frac{S_i - (l_i + 1) \mu_i}{\sum_{n=1}^{M} (S_n - (l_n + 1) \mu_n)}, \quad t \in T_R. \quad (15)$$

Hence, with CAS, the magnitude of $p_i$ only depends on the decision variables $\{S_i\}$. By defining $p_i$ as in (15), (14) simplifies to

$$I^i_t = (l_i + 1) \mu_i + p_i U_t. \quad (16)$$

The rational of CAS rationing is that it attempts to keep the ratio of the projected net inventory at any retailer over the system-wide projected net inventory constant at any time. This happens both when there is sufficient depot material ($t \in T_S$) and when material rationing is needed ($t \in T_R$). Notice that when $\Delta$ equals 0 AS rationing is equivalent to CAS.

For the general definition of AS rationing, we have $2M+1$ decision variables which should be determined
to obtain the desired system performance: \( \{S_n\}_{n=1}^M, \{p_n\}_{n=1}^M \) and \( S_0 \). Since

\[
p_M = 1 - \sum_{n=1}^{M-1} p_n
\]

we in fact have \( 2M \) decision variables. Notice that when \( \Delta \) equals 0 the number of decision variables reduces to \( M + 1 \), since \( \{S_n\} \) are irrelevant. With CAS rationing the number of variables which need to be determined are reduced to \( M + 1: \{S_n\}_{n=1}^M \) and \( S_0 \).

In Diks and De Kok [22] it is argued that one should adapt (16) by using different allocation fractions depending on whether \( U_i \) is positive or negative,

\[
I_i := (l_i + 1) \mu_i + p_i(U_i)^+ - q_i(-U_i)^+,
\]

where the allocation fraction \( q_i \) is defined as a function of \( p_i \), i.e., \( q_i = f_i(p_i) \). Diks and De Kok show that this function \( f_i \) has to be monotonously decreasing in \( p_i \), and \( \sum_{n=1}^M f_n(p_n) = 1 \). In practice the retailers usually require high service levels. Therefore, most of the times \( U_i \) is positive, which implies that the impact of \( q_i \) on \( I_i \) is very small. Due to this, and the fact that (17) is rather cumbersome to work with, it is reasonable to use (14), instead of (17).

By defining the order-up-to levels by

\[
S_i := (l_i + 1) \mu_i + \frac{\sigma_i \sqrt{l_i + 1}}{\sum_{n=1}^M \sigma_n \sqrt{l_n + 1}} \sum_{n=1}^M (S_n - (l_n + 1) \mu_n),
\]

the allocation fractions are given by (15): \( p_i = \sigma_i \sqrt{l_i + 1} / \sum_{n=1}^M \sigma_n \sqrt{l_n + 1} \). From (10) and (16) it immediately follows that CAS rationing constitutes a generalization of FS rationing. Finally, we like to emphasize that both AS and CAS rationing are more general than FS rationing since it holds for any demand distribution and can be used for any service definition desired (not just the stock-out probability), while FS requires normally distributed demands and identical target stock-out probabilities at all retailers.

**Numerical example**

We give some numerical results (see [9]) for a divergent three-echelon inventory system. The inventory system under consideration has an upstream stockpoint supplying six stockpoints, and each of these six stockpoints supplies a group of four stockpoints (see Fig. 3). The lead time of the most upstream stockpoint equals 5, the lead times of the stockpoints 1 to 6 all equal 2, and all other lead times (i.e., the lead times of the stockpoints in one of the six groups) equal 1. The demand characteristics of the stockpoints within one group are identical. Fig. 3 depicts the mean, \( \mu \), say, and the squared coefficient of variation, \( \sigma^2 \), say, of the customer demand at an end-stockpoint in every group (G1 to G6). The target fill rate \( \beta \) of every end-stockpoint is identical. We considered the following three values of \( \beta \): 0.9, 0.95 and 0.99. To control this inventory system we use the CAS rationing policy in every stockpoint. The control parameters (i.e., the allocation fractions and the order-up-to levels) are obtained by the method developed in [19] to ensure these target service levels \( \beta \). This method can only be used if \( \Delta_i \) is known in advance, where \( \Delta_i \) equals the order-up-to level of the stockpoint \( i \) minus the sum of all the order-up-to levels of the successors (cf. \( \Delta \) in Section 4.1.1). We assume that \( \Delta_2 = \Delta_4 = \Delta_5 = \Delta_6 = 0 \) and \( \Delta_1 \). The values of \( \Delta_0 \) and \( \Delta_1 \) are varied.

Table 3 depicts the results obtained from simulation of the model (using the control parameters obtained from the analysis). It yields insight in the performance of the CAS rationing policy and the impact of the imbalance on the attained external fill rates. When \( \Delta_0 = \Delta_1 = 0 \) every intermediate stockpoint in the multi-echelon inventory system does not hold any stock. This results in a considerable amount of imbalance: \( \Omega_1 = 0.55, \Omega_2 = 0.42 \) and \( \Omega_6 = 0.35 \), where \( \Omega_i \) represents the fraction of periods in which imbalance occurs in stockpoint \( i \) (in Section 4.2 we further elaborate on this measure for imbalance). Due to this
imbalance the service levels experienced by the customer differs significantly from the target fill rates ($\beta_i^g$ denotes the estimated fill rate of an end-stockpoint in group $i$). A well-known remedy to diminish the imbalance is to keep some stock back in the chain. In the model this corresponds with increasing one or more of $A_i$. Only raising $A_2$ and $A_3$ hardly has any effect. However, if at the same time $A_0$ is increased the gap between the attained fill rates and the target fill rates decrease considerably (see $\beta_i^g$ and $\beta_i^g$). $A_0$ is varied as 0, 14000 and 16000, which results in a mean stock of approximately 0, 0.2 and 0.8 period demands. Analogously $A_i$ is varied as 0 and 750, which results in a mean stock of approximately 0 and 0.4 period demands. The simulation shows the validity of the analysis for $N$-echelon model in [19] in case imbalance is sufficiently low. Table 3 also reveals that when some stock is kept in intermediate stockpoints the expected stock in the total supply chain increases with approximately 5%. Extensive numerical experiments applying the method of De Kok, Lagodimos and Seidel [20] and some discrete event simulation for validation purposes revealed that in most cases cost-optimal policies under service level constraint imply low stocks at intermediate stages.

4.1.3. Priority rationing

Lagodimos [38] introduces Priority rationing which closely resembles rationing policies used in practice. This policy produces a list of retailers and rations the available material so as to satisfy them in the sequence they were listed. Two particular rules for determining the allocation list are considered:

- RAN: assigning priorities at random.
- MIN: assigning priorities in order of increasing retailer order size.

The RAN rule is more often used in practice. The MIN rule is a natural extension of a rule suggested by Baker [8], and has the property of minimizing the number of unsatisfied retailers. Priority rationing can be considered as a pull policy [60], since all rationing decisions are effectively determined by the orders released by individual stockpoints. Notice that both FS and AS rationing are push policies, since rationing decisions are based on the system-wide inventory.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$A_0$</th>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>$A_4$</th>
<th>$A_5$</th>
<th>$A_6$</th>
<th>$\beta_1^g$</th>
<th>$\beta_2^g$</th>
<th>$\beta_3^g$</th>
<th>$\beta_4^g$</th>
<th>$\beta_5^g$</th>
<th>Stock (x1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.16</td>
<td>0.22</td>
<td>0.34</td>
<td>0.868</td>
<td>0.888</td>
<td>0.878</td>
<td>0.890</td>
</tr>
<tr>
<td>0.90</td>
<td>14000</td>
<td>0</td>
<td>0.03</td>
<td>0.54</td>
<td>0.22</td>
<td>0.42</td>
<td>0.16</td>
<td>0.22</td>
<td>0.35</td>
<td>0.884</td>
<td>0.892</td>
<td>0.887</td>
<td>0.893</td>
</tr>
<tr>
<td>0.90</td>
<td>16000</td>
<td>0</td>
<td>0.00</td>
<td>0.54</td>
<td>0.22</td>
<td>0.41</td>
<td>0.15</td>
<td>0.23</td>
<td>0.35</td>
<td>0.884</td>
<td>0.892</td>
<td>0.887</td>
<td>0.893</td>
</tr>
<tr>
<td>0.90</td>
<td>0</td>
<td>750</td>
<td>0.09</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.16</td>
<td>0.22</td>
<td>0.34</td>
<td>0.868</td>
<td>0.888</td>
<td>0.875</td>
<td>0.890</td>
</tr>
<tr>
<td>0.90</td>
<td>14000</td>
<td>750</td>
<td>0.03</td>
<td>0.27</td>
<td>0.22</td>
<td>0.19</td>
<td>0.16</td>
<td>0.22</td>
<td>0.35</td>
<td>0.896</td>
<td>0.892</td>
<td>0.895</td>
<td>0.893</td>
</tr>
<tr>
<td>0.90</td>
<td>16000</td>
<td>750</td>
<td>0.00</td>
<td>0.22</td>
<td>0.22</td>
<td>0.15</td>
<td>0.15</td>
<td>0.22</td>
<td>0.35</td>
<td>0.896</td>
<td>0.892</td>
<td>0.894</td>
<td>0.893</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.17</td>
<td>0.22</td>
<td>0.34</td>
<td>0.908</td>
<td>0.936</td>
<td>0.921</td>
<td>0.937</td>
</tr>
<tr>
<td>0.95</td>
<td>14000</td>
<td>0</td>
<td>0.03</td>
<td>0.54</td>
<td>0.22</td>
<td>0.42</td>
<td>0.16</td>
<td>0.22</td>
<td>0.34</td>
<td>0.933</td>
<td>0.942</td>
<td>0.936</td>
<td>0.942</td>
</tr>
<tr>
<td>0.95</td>
<td>16000</td>
<td>0</td>
<td>0.00</td>
<td>0.54</td>
<td>0.22</td>
<td>0.42</td>
<td>0.16</td>
<td>0.22</td>
<td>0.34</td>
<td>0.933</td>
<td>0.941</td>
<td>0.936</td>
<td>0.942</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>750</td>
<td>0.08</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.17</td>
<td>0.22</td>
<td>0.34</td>
<td>0.917</td>
<td>0.935</td>
<td>0.921</td>
<td>0.936</td>
</tr>
<tr>
<td>0.95</td>
<td>14000</td>
<td>750</td>
<td>0.03</td>
<td>0.27</td>
<td>0.22</td>
<td>0.18</td>
<td>0.16</td>
<td>0.22</td>
<td>0.35</td>
<td>0.947</td>
<td>0.942</td>
<td>0.946</td>
<td>0.943</td>
</tr>
<tr>
<td>0.95</td>
<td>16000</td>
<td>750</td>
<td>0.00</td>
<td>0.22</td>
<td>0.22</td>
<td>0.15</td>
<td>0.15</td>
<td>0.22</td>
<td>0.35</td>
<td>0.947</td>
<td>0.941</td>
<td>0.946</td>
<td>0.943</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.17</td>
<td>0.22</td>
<td>0.34</td>
<td>0.963</td>
<td>0.980</td>
<td>0.970</td>
<td>0.980</td>
</tr>
<tr>
<td>0.99</td>
<td>14000</td>
<td>0</td>
<td>0.03</td>
<td>0.54</td>
<td>0.23</td>
<td>0.42</td>
<td>0.17</td>
<td>0.22</td>
<td>0.35</td>
<td>0.981</td>
<td>0.985</td>
<td>0.982</td>
<td>0.986</td>
</tr>
<tr>
<td>0.99</td>
<td>16000</td>
<td>0</td>
<td>0.00</td>
<td>0.54</td>
<td>0.22</td>
<td>0.42</td>
<td>0.15</td>
<td>0.23</td>
<td>0.35</td>
<td>0.981</td>
<td>0.985</td>
<td>0.982</td>
<td>0.986</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>750</td>
<td>0.08</td>
<td>0.55</td>
<td>0.23</td>
<td>0.42</td>
<td>0.17</td>
<td>0.22</td>
<td>0.34</td>
<td>0.964</td>
<td>0.981</td>
<td>0.974</td>
<td>0.982</td>
</tr>
<tr>
<td>0.99</td>
<td>14000</td>
<td>750</td>
<td>0.03</td>
<td>0.27</td>
<td>0.22</td>
<td>0.18</td>
<td>0.16</td>
<td>0.22</td>
<td>0.35</td>
<td>0.989</td>
<td>0.986</td>
<td>0.989</td>
<td>0.986</td>
</tr>
<tr>
<td>0.99</td>
<td>16000</td>
<td>750</td>
<td>0.00</td>
<td>0.22</td>
<td>0.22</td>
<td>0.15</td>
<td>0.15</td>
<td>0.22</td>
<td>0.35</td>
<td>0.989</td>
<td>0.985</td>
<td>0.989</td>
<td>0.986</td>
</tr>
</tbody>
</table>
As in other rationing policies, it is extremely hard to model exactly all individual retailers inventories. Therefore the availability assumption is introduced, which assumes that in the event of material rationing, at most one retailer will not be completely satisfied (the last in the allocation list). From this assumption it follows that

\[ I_t^i = S_l - 1\{\alpha_l\}(D_{t-l_0,t} - \Delta)^+, \quad (18) \]

where \( \alpha \) is the index of the last retailer in the allocation list. The difference with CAS rationing becomes clear, if we rewrite (14)

\[ I_t^i = S_l - \alpha_l(D_{t-l_0,t} - \Delta)^+. \quad (19) \]

If during material rationing there is a shortage \( D_{t-l_0,t} - \Delta \), then in CAS rationing it is divided among the retailers, while in Priority rationing it is completely allocated to retailer \( \alpha \).

When in Priority rationing the RAN rule is used it is obvious that \( \alpha \sim U[1, M] \). This also holds when priority is given according to the MIN rule, provided that \( D_{t,t+1} \) is identically distributed for all \( i \).

### 4.1.4. Extensions and other rationing rules

Jönsson and Silver [36] analyze a divergent two-echelon model in which the depot has two shipping opportunities to ship products to the retailers, until the next depot replenishment. At the start of a replenishment cycle part of the depot stock is allocated and rationed among the retailers. At the second shipping opportunity the stock retained at the depot is allocated to the retailers so as to maximize the customer service \( (\gamma) \) until the time of the next replenishment. The applicability of this model in real-world situation is limited, since the approach is limited to two-echelon models with identical retailers, negligible lead times and normal distributed demand processes.

Verrijdt and De Kok [72] extends the divergent two-echelon model under CAS rationing to the more general divergent \( N \)-echelon model. In their paper intermediate stockpoints are not allowed to hold any stock. Later, De Kok [19] derived the results where the intermediate stockpoints are also allowed to hold stock. Recently, Verrijdt and De Kok [72] introduced the Balanced Stock rationing policy. He explains this policy by discussing a divergent two-echelon system where the depot is not allowed to hold stock. This policy rations the stock as in (19). However, the allocation fractions \( \{\alpha_l\} \) are determined such that the (approximate) mean imbalance is minimized. Next, the order-up-to levels \( \{S_l\} \) are determined such that every retailer attains his pre-determined target fill rate \( \beta_i \). Diks and De Kok [23] also use the rationing rule as defined in (19). They determine \( (R, S_l) \) policies using rationing rule (19), such that every retailer attains a pre-determined non-stock-out probability \( \alpha_i \) and a cost-function (based on the expected holding and penalty costs) is minimized. Their approach has been developed for divergent \( N \)-echelon inventory systems, where every intermediate stockpoint is allowed to hold stock.

So far the duration of the review period at the depot and at the retailers are identical. In Jackson [35] this assumption is relaxed, by assuming that the duration of the review period at the depot equals a multiple of the review period at the retailers. Without loss of generality we fix the review period of the retailers to one period, and \( m \in \mathbb{N} \) denotes the duration of the review period at the depot \( (m \notin \mathbb{N} \) results in a non-stationary demand process at the depot). When at time \( t = -l_0 \) the depot reviews its echelon inventory position, and places an order at the outside supplier. This order arrives at time \( t = 0 \). This marks the beginning of an order cycle at the depot, which ends at time \( m \) (just before the next arrival). Let \( \rho \) denote the time at which the depot runs out, or \( m \), whichever comes first

\[ \rho := \min(m, \inf\{t \mid D_{t-l_0,t} > \Delta\}). \]

This \( \rho \) is referred to as the pooled-risk period, since it can be thought of as the length of time within a cycle that risk is pooled. If \( \rho < m \) material rationing is necessary, otherwise no rationing is needed during this cycle. For \( t < \rho \) every retailer \( i \) is able to raise the inventory position to \( S_l \). At time \( t = \rho \) the depot stock has to be allocated over the retailers. Notice that after rationing the depot does not hold any stock on hand. The only time during the cycle that the rationing policy is effectively used is at time \( \rho \), since for \( \rho + 1 \leq t < m \) we have

\[ I_t^i = I_{t-1}^i - D_{t-1,t}. \]

As a consequence, only for \( t = \rho < m \) imbalance may occur. At this time \( \rho \) we could use any rationing policy (e.g. CAS), but Jackson [35] developed the run-out...
Allocation rule to determine \( I^*_p \). This rule minimizes the total amount of holding and penalty costs under the condition that

\[
\sum_{n=1}^{M} I^*_p = S_i - D_{i-1,p} \quad \text{and there is no imbalance, i.e.,}
\]

\[
I^*_p \geq S_i - D_{i-1,p}.
\]

To our knowledge little is known about divergent multi-echelon models which incorporate lot-sizing. \cite{66} is one of the few papers which deals with lot-sizing, since it analyzes a divergent two-echelon system where every stockpoint is controlled by an echelon stock \((R,s,nQ)\) policy. All retailers are identical and the control parameters are based on an \( \alpha \)-service level and normal demand.

### 4.1.5. Service levels

This section shows how to determine the three service level measures, introduced in Section 2.4, using rationing rule (19). We distinguish between the internal service (at the depot) and the external customer service:

#### The non-stock-out probability (\( \alpha \))

The rationing policies addressed enable us to determine both a tractable expression for the internal and external non-stock-out probability. The internal non-stock-out probability denotes the probability that the depot can satisfy all retail orders at once. It can be shown that this equals

\[
\Pr\{D_{t,t+l} \leq \Delta\}.
\]

The non-stock-out probability experienced by a customer of retailer \( i \) equals

\[
\alpha_i = \Pr\{I^*_i - D_{t,t+l} \geq 0\}.
\]

If every retailer requires an \( \alpha \)-service level, then for the determination of the control parameters of a rationing policy we refer to \cite{10,18,20,38,68,72,73}.

#### The fill rate (\( \beta \))

The internal fill rate at the depot, i.e., the fraction of the (retail) demand immediately delivered from the stock on hand, equals

\[
\frac{\sum_{n=1}^{M} \mu_n}{\sum_{n=1}^{M} (\mu_n + \mathbb{E}\{S_n - I^*_n\})}.
\]

Unlike the \( \alpha \) service level definition, the internal fill rate does not give any information about the internal fill rates experienced by the retailers. We can determine this by computing the fill rate issued at the depot to a particular retailer \( i \). This equals

\[
\frac{\mu_i}{\mu_i + \mathbb{E}\{S_i - I^*_i\}}.
\]

Much more important is the external fill rate of every retailer. For retailer \( i \) this fill rate yields

\[
\beta_i = 1 - \frac{\mathbb{E}\{(D_{t,t+l}^i - I^*_i)^+ - (D_{t,t+l}^i - I^*_i)^-\}}{\mu_i}.
\]

In order to numerically compute this fill rate we need to determine, among others, the expression \( \mathbb{E}\{(D_{t,t+l}^i - I^*_i)^+ - (D_{t,t+l}^i - I^*_i)^-\} \). If we use the rationing rule (19) rewriting this expression yields \( \mathbb{E}\{D_{t,t+l}^i + p_i(D_{t-1,l} - \Delta)^+ - S_i\}^+ \), and to determine its distribution is cumbersome. For multi-echelon systems with more than two stages, the expression involved becomes even less tractable, and therefore we use an approximate procedure, which is developed in \cite{54} (also see \cite{20}). Basically, this procedure fits a mixed Erlang distribution on the first two moments (cf. Tijms \cite{65}) of the random variable in each \((\cdot)^+\)-expression, beginning from the inside. Van Houtum and Zijm \cite{71} tested this approximation procedure extensively, and showed that it performs well. If every retailer requires a \( \beta \)-service level, then for the determination of the control parameters of a rationing policy we refer to \cite{10,18,20,38,68,72,73}.

#### The modified fill rate (\( \gamma \))

This service level is very similar to the fill rate definition. However, the modified fill rate issued at retailer \( i \) to the customers equals

\[
\gamma_i = 1 - \frac{\mathbb{E}(D_{t,t+l}^i - I^*_i)^+}{\mu_i}.
\]

Notice that \( \beta_i \) and \( \gamma_i \) are similar service measures. Specifically, if retailer \( i \) requires a high target service level, then usually the net stock of this retailer just after the arrival of a replenishment order is positive. Hence, \( \mathbb{E}(D_{t,t+l}^i - I^*_i)^+ \) is small, and thus \( \beta_i \) is approximately equal to \( \gamma_i \). If every retailer requires a \( \gamma \)-service level, then for the determination of the control parameters of a rationing policy we refer to Lagodimos \cite{38}.
4.2. Balance assumption

As already mentioned the FS policy rations the available material so as to maintain all end-stockpoints with the same stock-out probability. The phenomenon of not being able to achieve this is referred to as imbalance [26]. Verrijdt and De Kok [73] generalizes this definition by defining imbalance as the phenomenon where the rationing policy of an intermediate stockpoint allocates a negative quantity to at least one of its successors. In order to quantify the impact of the imbalance on the realized service levels, we need some analytical measure of imbalance. We focus on the probability that the depot allocates a negative quantity to one of the retailers,

$$\Omega := \Pr\{\exists i \in \{1, \ldots, M\} : q_i < 0\},$$

where $q_i$ is the amount allocated to retailer $i$ at time $t$. By definition

$$q_i^t = I_i^t - \bar{I}_i^t,$$

where $\bar{I}_i^t$ is the inventory position of retailer $i$ just before rationing at time $t$, thus $\bar{I}_i^t = I_{i-1}^t - D_{i-1,t}^t$. It is extremely difficult to derive an expression for the measure $\Omega$. Hence, Verrijdt and De Kok [73] use a surrogate measure:

$$\Omega_i^t := \Pr\{q_i^t < 0\} \quad \text{for } i = 1, \ldots, M.$$

Van der Heijden [68] suggested another way to measure the amount of imbalance,

$$\Omega_i^t := \mathbb{E}(q_i^t)^+ \quad \text{for } i = 1, \ldots, M.$$

In order to get a tractable expression for $q_i^t$ it is common [68,73] to assume that the previous allocation did not face any imbalance, since then $I_{i-1}^t$ is given by the rationing policy (e.g. (17), (10) or (18)). Under the aforementioned condition we obtain:

$$q_i^t = I_i^t - (I_{i-1}^t - D_{i-1,t}^t).$$

Then the total system imbalance is defined as $\hat{\Omega} := \sum_{n=1}^M \hat{\Omega}_i$. When imbalance occurs this means that one or more retailers have excess inventories. Therefore Van der Heijden [68] tries to determine the control parameters of the replenishment policy in order to minimize the imbalance, by minimizing $\hat{\Omega}$.

In the literature [18,26,38,73] it is very common to assume that the imbalance does not have a large impact on the results obtained from the analysis, e.g. the attained service levels at the end-stockpoints. This assumption is called the balance assumption [70]. For the validity of this assumption we refer to [66,73,75].

5. Conclusions and further research

In this paper we have shown that in the last two decades considerable progress has been made in the analysis of real-world supply chains. We focused on the analysis of control policies with respect to external customer service and supply costs. The distinction between installation stock policies and echelon stock policies revealed that the complexity of the analysis is concentrated at different aspects for the two types of policies. For installation stock policies the major difficulties are the determination of the demand process at upstream stages and the determination of the delay time characteristics. Most of the literature on installation stock policies circumvents this problem by making particular assumptions, of which the assumption of pure Poisson demand together with identical lot sizes for all downstream stockpoints are the most important ones. For echelon stock policies the major problems are the imbalance problem and the analysis of divergent systems with lot sizing. Furthermore hardly any results are available on continuous review echelon stock policies. Since our aim was to identify practically useful approximation procedures for the analysis of multi-echelon divergent systems with service level constraints we give below a summary of the state-of-the art. In Table 4 we classify multi-echelon divergent systems according to a number of characteristics. For each combination of characteristics we give the paper(s) that we consider to be most applicable to real-world situations up-to-now. This is a subjective selection based on one of the authors extensive knowledge of real-world supply chains. E.g. this implies that approximation schemes for identical retailers are considered non-applicable. On the other hand, approximation schemes based on the assumption of identical lot sizes are considered to be applicable, although to a limited extent. From Table 4 it is easy to identify the white spots for which further research is required. For installation stock policies we need approximation schemes for heterogeneous systems with lot sizing. Furthermore more work is required to find (approx-
Table 4
Overview of the most recent work in the different areas (italics means that models can be extended to N-echelon systems)

<table>
<thead>
<tr>
<th>Stock Policy</th>
<th>Periodic</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installation Order-up-to $s = S$</td>
<td>van der Heijden (1989)</td>
<td>Sherbrooke (1986)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Axsäter (1990)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Axsäter et al. (1994)</td>
</tr>
<tr>
<td></td>
<td>$s &lt; S$</td>
<td>Schneider et al. (1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Axsäter et al. (1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>van Donselaar (1990)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>$s &lt; S$</td>
<td></td>
</tr>
<tr>
<td>Batch order</td>
<td>Chen &amp; Zheng (1994)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Chen &amp; Zheng (1994)&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> Poisson processes.  <sup>b</sup> Stuttering Poisson processes.  <sup>c</sup> Identical retailers.  <sup>d</sup> Serial system.

Approximately cost-optimal policies. The work of Schneider, Rinks and Kelle [52] can be seen as a first step. Quite interesting is the paper by Axsäter, Forsberg and Zhang [4]. The approach presented there more or less implies that the performance of arbitrary divergent networks can be determined from "equivalent" networks with pure Poisson demand and $(S - 1, S)$ policies. We feel that further verification of this claim is required. For multi-echelon divergent systems further research is required into models with lot sizing. The comparison studies by Axsäter and Rosling [6] provide interesting material to start performance comparisons between echelon stock policies, DRP/MRP policies and DRP/MRP policies with order release restrictions. Finally we advocate the establishment of a set of multi-echelon system instances for benchmarking purposes. Today many papers report numerical results which are impossible to replicate for verification, validation and comparison purposes. In other cases it is not clear whether results presented are representative and practically relevant. It is our intention to set up such a benchmark set.

References


[34] Iglehart, D., “Optimality of (s, S) policies in the infinite horizon dynamic inventory problem”, Management Science 9 (1963) 259–267.


[50] Schneider, H., “Methods for determining the reorder point of an (s, S) ordering policy when a service-level is specified”, Operational Research Quarterly 29 (1978) 1181–1193.


